A note on the generalized degrees of freedom under the L_1 loss function

By: Xiaoli Gao, Yixin Fang

Gao, X.L. and Fang, Y.X. (2011). A note on the generalized degrees of freedom under the L1 loss function. *Journal of Statistical Planning and Inference*, 141(2), 677-686. doi:10.1016/j.jspi.2010.07.006

Made available courtesy of Elsevier: http://dx.doi.org/10.1016/j.jspi.2010.07.006

Abstract:

Generalized degrees of freedom measure the complexity of a modeling procedure; a modeling procedure is a combination of model selection and model fitting. In this manuscript, we consider two definitions of generalized degrees of freedom for a modeling procedure under the L_1 loss function, and investigate the connections between those two definitions. We also propose the extended Akaike information criterion, the adaptive model selection, and the extended generalized cross-validation under the L_1 loss function. Finally, we extend the results to M-estimation.

Keywords: Adaptive model selection | Covariance penalty | Degrees of freedom | Generalized cross-validation | Least absolute deviations | Modeling procedure

Article:

1. Introduction

In the literature of model selection, modeling procedures are usually assessed according to prediction errors. Based on training data $y=(y_1,\ldots,y_2)'$, a modeling procedure is constructed and an estimator $\hat{\mu}$ is produced, and then the prediction error $L(y^0,\hat{\mu})$measures how well the modeling procedure predicts the future data $y^0=(y_1^0,\ldots,y_2^0)'$. Here L is a loss function and $\mathbf{y}^0$ is independently generated from the same mechanism that generated $\mathbf{y}$. See for example, Efron (1986, 2004).

It is important to notice that a modeling procedure is different from a model fitting. As pointed out by Ye (1998), a modeling procedure M is a combination process of both model selection and model fitting mapped by

$$\mathcal{M}: \mathbf{y} \overset{\text{selection}}{\Longrightarrow} \text{selected model } M \overset{\text{fitting}}{\Longrightarrow} \hat{\boldsymbol{\mu}}.$$

The degrees of freedom (DF) measure the complexity of a model fitted a priori. Ye (1998) introduced the concept of generalized degrees of freedom (GDF) to measure the complexity of a modeling procedure. In the frame of the L_2 loss function and Normal assumption, Ye (1998) interpreted the GDF as "the sum of sensitivity of each fitted value to the perturbation in the corresponding observed value."

As we know, the least squares (LS) regression may fail to produce a reliable estimator when the dataset subjects to heavy-tailed errors, and the least absolute deviations (LAD) estimator enjoys many good properties such as robustness to the outliers. Without any theoretic justification, some recent work used the concept of GDF directly in the selection of the optimal regularization parameter. For example, the GDF was used directly by Nychka et al. (1995), Yuan (2006), and Li et al. (2007) in quantile smooth splines, and Li and Zhu (2008) in L_1-norm quantile regression. In their work, a modeling procedure is amount to choosing an optimal regularization parameter, say λ_{opt}. Therefore, model selection procedures under the L_1 loss function deserve further investigation. The main goal of this manuscript is to justify the rationale of two definitions of GDF under the L_1 loss function: one based on the sensitivity of perturbation, and the other based on the covariance penalty. Furthermore, the connections between these two definitions, the application of GDF in correcting the bias, and the extension to M-estimation are discussed.

The remaining of the manuscript is organized as follows. In Section 2, we review some modeling procedures under the L_1 loss function. In Section 3, we give two definitions of GDF under the L_1 loss function, and investigate their connections. In Section 4, we propose the extended Akaike information criterion (EAICR), the adaptive model selection, and the extended generalized cross-validation (EGCV) under L_1 loss function. We extend the results to M-estimation in Section 5. In Section 6, we conduct some numerical studies. We conclude the paper with some discussions in Section 7. Some technical proofs are given in Appendix.

2. Review on some modeling procedures

Consider the following linear model:

(1) $y_i=x'_i\beta+\varepsilon_i, i=1,\ldots,n,$

where y_i's are independent observations, $\mathbf{x}_i$'s are fixed $p\times1$ predictors, β is a $p\times1$ coefficient vector, and ε_i's are assumed to be independently and identically distributed, with median 0 and a continuous, positive density $f(x)$ in a neighborhood of 0. For convenience, let the first component of $\mathbf{x}_i$ be one and incorporate the intercept into β. Let α be any subset of $\{1,\ldots, p\}$ with $d\alpha$ distinct elements. Let $\beta\alpha$ and $x_{\alpha i}$ be the corresponding $d\alpha$-vectors indexed by set α. Let $M\alpha$ be the sub-model, $y_i=x'_{\alpha i}\beta\alpha+\varepsilon_i$. Under the L_1 loss function, if a model $M\alpha$ is fitted a priori, the LAD estimator of $\beta\alpha$ is defined as

(2) $\hat{\beta}_a = \arg\min_{\gamma_a} \Sigma_{i=1}^{n} |y_i - x'\alpha i\gamma\alpha|.$

It is well-known that if εi follows a double exponential distribution DE(0,σ),$\hat{\beta}$α is actually the MLE of βα, and the MLE of σ is

(3) $\hat{\sigma}_\alpha = \sum_{i=1}^{n} \left| y_i - X'_{\alpha_i} \hat{\beta}\alpha \right| / n.$

Naturally, $\hat{\mu}_i$(=x′αi$\hat{\beta}_\alpha$) provides an estimator for μi(=x′iβ) under model Mα. If Mα is given a priori, then its complexity is measured by the DF, which simply equals dα.

Any criterion can be treated as a modeling procedure when used to choose a model to fit the data. The most famous one is the Akaike information criterion (AIC, Akaike, 1973). For any model Mα, the AIC score is defined as

AIC(α)=−2Lα+2dα,

where Lα is the log-likelihood under model Mα. The AIC modeling procedure is to choose the optimal model that minimizes AIC(α). When εi follows a double exponential distribution,

AIC(α)=cn+2nlog($\sum_{i=1}^{n} \left| y_i - x'_{\alpha i} \hat{\beta}_\alpha \right|$)+2dα,

where c_n=2n log(2/n)+2n. Later work generalizes AIC to the form of −2Lα+penalty, where the penalty term penalizes the model complexity. For examples, the penalty term in BIC (Schwarz, 1978) is dαlog(n), in HQIC (Hannan and Quinn, 1979) is 2dαloglog(n), and in cAIC (Hurvich and Tsai, 1989) is 2(dα+1)(dα+2)/(n−dα−2). Especially, Ronchetti (1985) proposed a robust model selection criterion (AICR). That is, for a given λ>0,

(4)AICR(α;ρ,λ)=2 $\sum_{i=1}^{n} p\left(\left(y_{i-x'\alpha_i} \hat{\beta}_\alpha \right) / \hat{\sigma} \right) + \lambda d_\alpha,$

where $\hat{\beta}_\alpha$ is an M-estimator such that $\sum_{i=1}^{n} \Psi \left(\left(y_{i-x'\alpha i} \hat{\beta}_\alpha \right) / \hat{\sigma} \right) x_{\alpha_i}$=0 with ψ(u)=dρ(u)/du and $\hat{\sigma}$ is a robust estimator of σ. The LAD estimator is an M-estimator by choosing ρ(u)=|u|. Thus for λ=2, AICR under the L_1 loss function can be rewritten as

(5)AICR(α)=∑i=1n|yi−x′αi$\hat{\beta}$α|+dα$\hat{\sigma}$.

3. GDF under L_1 loss function

In this section we give two definitions of GDF under the L_1 loss function for any modeling procedure, where a modeling procedure includes selecting an optimal sub-model Mα^, finding the LAD estimator $\hat{\beta}_{\hat{\alpha}}$, and then obtaining the fitted values $\hat{\mu}_i$=x′$\hat{\alpha}_i\hat{\beta}_{\hat{\alpha}}$.

3.1. GDF: on the sensitivity to perturbation

Under the Normal assumption and L_2 loss function, Ye (1998) defined the GDF as the sum of the sensitivity of each fitted value to perturbation in the corresponding value. We extend the definition to the L_1 loss function.

Definition 1

The GDF of a modeling procedure M under the L_1 loss function is defined as

$$D(M)=\sum_{i=1}^{n} E\left[\frac{\partial \hat{\mu}_i(y)}{\partial y_i}\right]=\lim_{\delta\to 0}\sum_{i=1}^{n} E\left[\frac{\hat{\mu}_i\left(y+\delta_{e_i}\right)-\hat{\mu}_i(y)}{\delta}\right],$$

where $\hat{\mu}_i(y)$ is the fitted value of y_i based on data $\mathbf{y}$ through the modeling procedure M, and $\mathbf{e}_i$ is the ith column of the $n\times n$ identity matrix.

We can use the Monto Carlo algorithm proposed by Ye (1998) to estimate D(M). The LAD estimator $\hat{\beta}$ may not be unique, but the estimator $\hat{\mu}$ is unique. As in Li et al. (2007), Li and Zhu (2008), we assume that data (y_i,$\mathbf{x}_i$), …, (y_i, $\mathbf{x}_n$) are in general positions such that the LAD estimator is unique throughout the manuscript.

First and foremost, we should verify that the GDF in Definition 1 is consistent with the DF, the number of predictors in model Mα, if a model Mα is fitted a priori. Without this, any definition of GDF under the L_1 loss function seems risky. For a LAD estimator β^ in a given model M with d predictors, following the spirit of Li and Zhu (2008), we define the elbow set as Ey={1≤i≤n:yi−x′iβ^=0} and Nx as the set of $\mathbf{y}$ such that |Ey|>d for any given designed covariates {$\mathbf{x}_i$}'s, where |·| is the cardinal value. We have the following result.

Theorem 1

Assume that model Mα*is fitted a priori. For any* $y\in R^n\setminus N_x$, *we have*

$$\sum_{i=1}^{n}\frac{\partial \hat{\mu}(y)}{\partial y_i}=d\alpha.$$

In fact, $P(N_x)=0$ if y_i has a continuous density. Thus, $D(M)=E[\sum_{i=1}^{n}\partial\hat{\mu}_i/\partial y_i]=d\alpha$. That is, the GDF is consistent with the DF when Mα is fitted a priori. The proof of Theorem 1 is given in Appendix.

3.2. GDF: on the covariance penalty

Under the L_2 loss function, Efron (2004) used the covariance penalty to correct the bias while using the apparent error (err) to estimate the expected prediction error (Err). He considered a bigger class of loss functions, the q-class. However, the q-class does not include the L_1 loss function. To adopt the spirit of the covariance penalty to the L_1 loss function, we provide another definition of GDF under the L_1 loss function.

Definition 2

The GDF of a modeling procedure M under L_1 loss function is defined as

(6)D(M)=1σ∑i=1nEsgn(yi−μi)(μ^i(y)−μi)=1σ∑i=1nCov{sgn(εi),μ^i(y)},

where sgn(x)=1,0,−1 when x>0,=0,<0.

We can use the parametric Bootstrap (Efron and Tibshirani, 1993) to estimate the GDF in Definition 2. Let $\{\varepsilon_1^*,\ldots,\varepsilon_n^*\}$ be a random sample of size n from the empirical distribution of the observed residuals $\{\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_n\}$, $\mathbf{y}^*$ be a Bootstrap sample of $\{y_i^*=\hat{\mu}_i+\varepsilon_i^*,i=1,\ldots,n\}$, $\{\hat{\mu}_1(y^*),\ldots,\hat{\mu}_n(y^*)\}$ be the fitted values based on the Bootstrap sample, and $\mathbf{F}^*$ be the Bootstrap distribution of $\mathbf{y}^*$. Then, the covariance term in (6) can be estimated by $\Sigma_{i=1}^n \mathrm{Cov}_F^*\{\mathrm{sgn}(\varepsilon_i^*),\hat{\mu}_i(y^*)\}$.

Again, we should verify that the GDF in Definition 2 is consistent with the DF. In model (1), suppose $\Sigma_{i=1}^n x_i x'_i$ has a positive definite square root V_n and $\max_{1\le i\le n}|Vn-1xi|\to 0$ as $n\to\infty$. By the asymptotic Normality of the LAD estimator shown in Pollard (1991), we have

$$\frac{1}{\sigma}\Sigma_{i=1}^n\{\mathrm{sgn}(\varepsilon i)(\hat{\mu}_i(y)-\mu_i)\}=[\Sigma_{i=1}^n n\mathrm{sgn}(\varepsilon i)V_n^{-1}x_{\alpha i}]'[\Sigma_{i=1}^n \mathrm{sgn}(\varepsilon_i)V_n^{-1}x_{\alpha i}]+o_P(1)=d_\alpha+o_P(1).$$

Therefore, the GDF defined in Definition 2 is almost to DF when a model M_α is fitted a priori.

For a modeling procedure, say AICR, we first select a "best" sub-model Mα^, then fit Mα^ to the data and get a LAD estimator . Let $\mathbf{F}_0$ be the distribution of future data $\mathbf{y}^0$. Under the L_1 loss function,

(7) $\mathrm{Err}=E_{F0}\Sigma_{i=1}^n|y_i^0-x'\hat{\alpha}_i\hat{\beta}\hat{\alpha}|$

measures how well the modeling procedure predicts the future data y_i^0, and

(8) $\mathrm{err}=\Sigma_{i=1}^n|y_i-x'\hat{\alpha}_i\hat{\beta}\hat{\alpha}|$

measures only the goodness of fit. Since a more complex modeling procedure leads to a smaller err, we cannot use only err to compare the performances of modeling procedures. However, since Err reflects a trade-off between the goodness of fit and the complexity of a modeling procedure, we can use it to compare modeling procedures. The following result provides us a heuristic approximation to Err.

Theorem 2

Let Err and err be defined in (7) and (8). *We have*

(9) $E\{\mathrm{Err}\}=E\{\mathrm{err}\}+E\{\Sigma_{i=1}^n \mathrm{sgn}(y_i-\mu_i)(\hat{\mu}_i-\mu_i)\}+E[h_n(y)]$,

where $h_n(\mathbf{y})=o_P(1)$.

Similar to Efron (1986), we can interpret $\Sigma_{i=1}^n \mathrm{sgn}(y_i-\mu_i)(\hat{\mu}_i-\mu_i)$ in (9) as an "almost" optimism and σD(M) as the expected optimism. Thus, Theorem 2 provides an "almost" unbiased estimator of Err,

(10) $\Sigma_{i=1}^n|y_i-x'_{\alpha i}\hat{\beta}\alpha|+\sigma D(M)$.

3.3. Connections between two definitions

Here we investigate the connections between the above two definitions. Analogous to Lemma 1 in Stein (1981), we propose the following Lemma 1.

Lemma 1

Let Y be a random variable following DE(μ,σ). *For any function* g:R→R *with derivative* g˙ *and* E|g˙(Y)|<∞, *we have*

E[sgn(Y−μ)g(Y)]=σE[$\dot{g}$(Y)].

The proof of Lemma 1 is omitted since it is a special case of Lemma 3 in Section 5.

Theorem 3

GDFs defined in Definitions 1 and 2 *are equivalent when* εi *follows* DE(0,σ).

Theorem 3 is a direct result of Lemma 1; it builds a connection between Definitions 1 and 2 under the double exponential assumption.

4. Some applications of GDF under L_1 loss functions

4.1. Extended AICR

The AICR criterion defined in (5) can only be used for comparing models instead of modeling procedures. If we estimate the GDF, D(M), by D^(M), then (10) provides us an extended AICR criterion (EAICR) to compare modeling procedures under the L_1 loss,

(11) EAICR=$\Sigma_{i=1}^{n}$|yi−$\hat{\mu}$i|+$\widehat{D}$(M)$\hat{\sigma}_A$.

The comparison between EAICR and AICR is analogous to one between EAIC and AIC proposed in Ye (1998) under the L_2 loss function.

4.2. Adaptive model selection

As an application of the concept of GDF, we propose a version of adaptive model selection under the L_1 loss function. This is motivated by Shen and Ye (2002), where a version of adaptive model selection was proposed under the Normal assumption and L_2 loss function, and Shen et al. (2004), where a version of adaptive model selection was proposed under Exponential-Family distributions and Kullback–Leibler loss function.

Under the L_1 loss function, a robust model selection criterion (4) can be simplified as

(12) AICRλ=$\Sigma_{i=1}^{n}$|yi−$\hat{\mu}_i(M_\alpha)$|+λ$d_\alpha\hat{\sigma}$,

with respect to sub-models Mα. For a given λ, the corresponding model selection procedure is amount to choosing the optimal model M$\hat{\alpha}$(λ) to minimize AICRλ. Noting that for any given λ>0, (12) is actually a modeling procedure

$$\mathcal{M}_\lambda : \mathbf{y} \overset{\text{selection}}{\Longrightarrow} \text{selected model } M_{\hat{\alpha}}(\lambda) \overset{\text{fitting}}{\Longrightarrow} \hat{\boldsymbol{\mu}}(M_{\hat{\alpha}}(\lambda)).$$

Naturally, we attempt to choose an optimal λ, which has the “best” performance. If the GDF of Mλ, D(Mλ), is estimated by D^(Mλ), then the optimal λ^ is obtained by minimizing

(14) EAICR$_\lambda$= $\Sigma_{i=1}^{n}$|yi−$\hat{\mu}$i(M$\hat{\alpha}$(λ))|+D^(M$_\lambda$) $\hat{\sigma}_A$,

with respect to λ>0. Once a data-adaptive $\hat{\lambda}$ is obtained, we can select an optimal sub-model, M$\hat{\alpha}$ ($\hat{\lambda}$).

4.3. Extended generalized cross-validation

Under the L_2 loss function, the ordinary leave-one-out cross-validation (OCV) is a popular method (e.g., Stone, 1977) to provide an unbiased estimator of Err. Under the L_1 loss, it is natural for us to estimate Err of model Mα in (7) by

OCV$_{M\alpha}$=$\Sigma_{i=1}^{n}$|y$_i$−x′$_{\alpha i}\hat{\beta}_\alpha^{[i]}$|,

where $\hat{\beta}_\alpha^{[i]}$ is the LAD estimator of βα using data without subject *i*.

Following the spirit in Craven and Wahba (1979), we obtain the following leave-one-out lemma.

Lemma 2

For k and z, let $\hat{\beta}_\alpha$[k,z]*be the solution of*

$$\arg\min_{\gamma\alpha} |z-x'_{\alpha k}\gamma_\alpha|+\sum_{j\neq k}|yj-x'_{\alpha j}\gamma_\alpha|.$$

Then $\hat{\beta}$α[i,$\tilde{y}_i$]= $\hat{\beta}_\alpha^{[i]}$, *where* $\tilde{y}$i=x′$_{\alpha i}\hat{\beta}_\alpha^{[i]}$.

If we consider

y$_i$−x′$_{\alpha i}\hat{\beta}_\alpha^{[i]}$=(y$_i$−x′$_{\alpha i}\hat{\beta}$α)/(1−h*$_{\alpha i}$),

where h*$_{\alpha i}$=$\left(x'_{\alpha i}\hat{\beta}-x'_{\alpha i}\hat{\beta}^{[i]}\right)$/(y$_i$-x′α$i\hat{\beta}_\alpha^{[i]}$), then from Lemma 2, x′$_{\alpha i}\hat{\beta}_\alpha$[i,$\tilde{y}_i$]=x′$_{\alpha i}\hat{\beta}_\alpha^{[i]}$, where y˜i=x′αiβ^α[i]. Thus we have

h*$_{\alpha i}$=$\frac{x'\alpha i\hat{\beta}[i,y_i]-x'\alpha i\hat{\beta}_\alpha[i,\tilde{y}_i]}{yi-\tilde{y}i}$≐hαi=$\triangleq \frac{\partial\hat{\mu}_i}{\partial y_i}$.

Therefore, we can approximate OCV as

OCVM$_\alpha$≐$\Sigma_{i=1}^{n}$|yi−x′$_{\alpha i}\hat{\beta}_\alpha$|/(1−h$_{\alpha i}$).

Furthermore, if we replace hαi by E{$\Sigma_{j=1}^{n}$h$_{\alpha j}$}/n=D(M$_\alpha$)/n, we obtain EGCV defined by

(15) EGCV$_{M\alpha}$=$\Sigma_{i=1}^{n}$|yi−x′$_{\alpha i}\hat{\beta}_\alpha$|/(1−D(Mα)/n).

Noting that 1/(1−x)≐1+x when *x* is small. Thus, the EGCV is similar to the EAICR. EGCV is also consistent to GACV proposed by Yuan (2006) if we apply Definition 1 of GDF in Section 3.1. The above connection also exists when we are interested in comparing different modeling procedures. For example, we consider the adaptive model selection procedure in Section 3.2. For a given λ, Mλ is a modeling procedure, and the OCV method provides an unbiased estimator of the prediction error for Mλ. For given λ and *i*, let α^[i](λ) be the selected optimal model through

modeling procedure Mλ using data without subject *i*, and let $\hat{\mu}_{\lambda}^{[i]}$(=$x'\hat{\alpha}_{i}^{[i]}i\hat{\beta}_{\alpha}^{[i]}$) be the fitted value of y_i based on $M_{\hat{\alpha}[i]}$(λ). Then OCV is defined as

$OCV_{\lambda}=\Sigma_{i=1}^{n}|y_i-\hat{\mu}_{\lambda}^{[i]}|$.

Following similar arguments, for a modeling procedure M_{λ}, we define

$EGCV_{\lambda}=\Sigma_{i=1}^{n}|y_i-\hat{\mu}_{\lambda}|/(1-D(M\lambda)/n)$,

where $\hat{\mu}_{\lambda}$ is the fitted value of y_i using all the data, and $D(M_{\lambda})$ is the GDF of the modeling procedure M_{λ}. Thus $EGCV_{\lambda}$ is similar to $EAICR_{\lambda}$.

5. GDF in M-estimation

In this section, we extend concepts of the GDF to M-estimation. In linear model (1), an M-estimator of βα under any sub-model Mα is

(16) $\hat{\beta}_{\alpha}=\arg\min_{\gamma_{\alpha}} \Sigma_{i=1}^{n}\rho(y_i-x'_{\alpha i}\gamma_{\alpha})$,

where ρ(·) is a known function of errors with the influence function ϕ(u)(=dρ(u)/du). Also, $\hat{\mu}_i(=x'_{\alpha i}\hat{\beta}_{\alpha})$ provides us a robust estimator of $\mu_i(=x'_i\beta)$ under the model M_{α}. For any modeling procedure under M-estimation, say AICR in (4), we first select a "best" sub-model $M\hat{\alpha}$, and then fit the data to get an M-estimator $\hat{\beta}_{\hat{\alpha}}$.

Consider a generalized tp,q(μ,σ) distribution with the probability density

$f(x;\mu,\sigma,p,q)=c_{p,q}\sigma^{-1}(1+|x-\mu|^p/(q\sigma p))^{-q-1/p}, i=1,\ldots,n$,

where $c_{p,q}=(p/2)\, q^{-1/p}B^{-1}(1/p, q)$. We have the following result

Lemma 3

Let Y be a random variable following a generalized tp,q(μ,σ)*distribution. Let* g:R→R *have derivative* $g^{\cdot}$ *with* $E|g^{\cdot}(Y)|<\infty$. *Then*

$E[\psi(\frac{y-\mu}{\sigma})g(Y)]=\sigma E[g^{\cdot}(Y)]$,

where $\psi(u)=(p+1/q)|u|^{p-1}(1+|u|^p/q)^{-1}sgn(u)$.

Generalized tp,q(μ,σ) distribution is related to many distributions. For example, when q→∞, *Y* in Lemma 3 has a generalized error distribution with probability density

$f(x;\mu,\sigma,1,\infty)=p/(2\sigma\Gamma(1/p))\exp\{-|x-\mu|^p/\sigma^p\}$.

Thus, Lemma 1 in Section 3 and Stein's lemma are two special cases of Lemma 3 when p=1,q→∞ and p=2,q→∞.

When *Y* follows generalized $t_{p,q}(\mu,\sigma)$, an M-estimator is obtained by choosing $\rho(u)=(p+1/q)\log(1+|u|^p/q)$. In the light of Lemma 3, we provide two definitions of the GDF for any modeling procedure M in M-estimation.

Definition 3

Suppose y_i in linear model (1) follows generalized $t_{p,q(}\mu,\sigma)$. The GDF in M-estimation is defined as

$\Sigma_{i=1}^{n}$E[$\frac{\partial\hat{\mu}_i(y)}{\partial\mu_i}$]=$\frac{1}{\sigma}$∑i=1nE[ψ($\frac{y_i-\mu_i}{\sigma}$)($\hat{\mu}$(y)−μi)]=$\frac{1}{\sigma}$

$\Sigma_{i=1}^{n}$Cov{ψ($\frac{y_i-\mu_i}{\sigma}$),$\hat{\mu}_i$(y)},

where $\psi(u)=(p+1/q)p|u|^{p-1}(1+|u|^p/q)^{-1}sgn(u)$.

Definition 4

Suppose y_i in linear model (1) satisfies $E[\psi((yi-\mu_i)/\sigma)g(y_i)]=E[g^{\cdot}(y_i)]$. The GDF in M-estimation with influence function ψ is defined as

$\Sigma_{i=1}^{n}$E[$\frac{\partial\hat{\mu}_i(y)}{\partial\mu_i}$]=$\frac{1}{\sigma}$∑i=1nE[ψ($\frac{y_i-\mu_i}{\sigma}$)($\hat{\mu}$(y)−μi)]=$\frac{1}{\sigma}$

$\Sigma_{i=1}^{n}$Cov{ψ($\frac{y_i-\mu_i}{\sigma}$),$\hat{\mu}_i$(y)},

Thus, those results of the GDF can be extended to M-estimation. For example, we can use the defined GDF to correct the bias in M-estimation.

6. Numerical studies

In this section, we conduct simulation studies under the L_1 loss function to investigate: (1) the performance of GDF when a model is given a priori; (2) the performance of GDF for modeling procedures.

Consider linear models (1) with sample size n=50 for all the simulation studies. We generated the p-vector predictors $\mathbf{x}_i$'s independently from a multivariate Normal distribution with mean $\mathbf{0}_p$ and covariance matrix $\Sigma_x(=(1-\rho)I_p+\rho 1_p 1'_p)$; random errors εi's independently from three different types of distributions: (i) standard double exponential distribution DE(0,1); (ii) standard Normal distribution N(0,1); and (iii) standardized t-distribution t(3). For simplicity, we only consider the zero intercept. We denote aq0p−q as the vector with the first q components being a and the remaining being 0. Each simulation setting is repeated by 500 times.

We examine Theorem 1 under three residual distributions by both Examples 1 and 2. Meanwhile, we compare the Monto Carlo procedure in Definition 1 with the Bootstrap procedure in Definition 2 for computing the GDF in Example 1.

Example 1

Assume that p=8 and ρ=0.5 or 0.8. The true parameter $\beta_0=(3,1.5,0,0,2,0,0,0)'$. Five models, the truly most parsimonious model $M_0=\{1, 2, 5\}$, $M_1=M_0\cup\{3\}$,$M_2=M_1\cup\{4\}$,$M_3=M_2\cup\{6\}$, and the full model M_4 are considered. GDFs are evaluated when each of the above five models is given a priori.

We compute the GDF using both the Monto Carlo procedure in Definition 1 and the Bootstrap procedure in Definition 2. We generate B=500 Bootstrap samples when the parametric Bootstrap procedure is applied.

Results of the averages and standard deviations of GDF from 500 iterations for Example 1. Those simulation results support Theorem 1 for all three residual distributions; that is, if a model is given a prior, the GDF in Definition 1 is equal to the number of predictors in the model. The Bootstrap procedure also provides us a good estimator of the GDF, especially when the given model is relatively close to the true one. However, the Bootstrap method generates more bias than the Monto Carlo method, especially when the given model is a full model. Although we only justify the connections of the two definitions under the assumption of the double exponential distribution, our simulation studies demonstrate the connections of these two definitions under all three distributions. Table 1 also shows us that the strength of the correlation among predictors does not affect the simulation results significantly.

Example 2

Assume that p=20, and the true model $\beta_0=2_5 0_{15}$. The GDFs are evaluated for six models fitted a priori: under-fitted model $M_1=\{1\}$, correct model $M_{5,1}=\{1,\ldots,5\}$, wrong model $M_{5,2}=\{3,\ldots,7\}$, over-fitted model $M_{10,1}=\{1,\ldots,10\}$, noise-predictor-only model $M_{10,2}=\{6,\ldots,15\}$ and full model $M_{20}=\{1,\ldots,20\}$.

Table 1. GDF of models given a priori in Example 1, $\beta_0=(3,1.5,0,0,2,0,0,0)'$.

		Model				
		$\boldsymbol{M_0}$	$\boldsymbol{M_1}$	$\boldsymbol{M_2}$	$\boldsymbol{M_3}$	$\boldsymbol{M_4}$
	DF	**3**	**4**	**5**	**6**	**8**
ρ=0.8						
DE(0,1)	GDF	2.99 (0.31)	3.98 (0.35)	5.00 (0.39)	5.95 (0.43)	7.96 (0.53)
	GDF_{BP}	3.11 (0.10)	4.04 (0.13)	4.52 (0.15)	5.70 (0.16)	7.66 (0.18)
N(0, 1)	GDF	3.00 (0.42)	4.00 (0.50)	5.00 (0.52)	6.06 (0.55)	8.07 (0.87)
	GDF_{BP}	3.10 (0.11)	4.00 (0.13)	5.05 (0.15)	6.19 (0.16)	7.84 (0.20)
t(3)	GDF	3.00 (0.42)	4.04 (0.51)	4.97 (0.65)	6.00 (0.41)	8.02 (0.40)
	GDF_{BP}	3.02 (0.13)	4.48 (0.15)	5.02 (0.15)	5.94 (0.16)	7.08 (0.19)
ρ=0.5						
DE(0,1)	GDF	3.00 (0.28)	3.99 (0.28)	5.02 (0.30)	6.03 (0.36)	8.00 (0.45)
	GDF_{BP}	2.97 (0.12)	4.02 (0.14)	4.92 (0.14)	6.04 (0.18)	7.27 (0.20)
N(0, 1)	GDF	3.00 (0.54)	4.00 (0.63)	5.00 (0.67)	5.98 (0.48)	8.05 (0.76)
	GDF_{BP}	3.01 (0.28)	4.17 (0.31)	4.60 (0.37)	5.57 (0.40)	7.67 (0.47)
t(3)	GDF	3.00 (0.31)	3.99 (0.34)	5.01 (0.31)	6.00 (0.32)	8.00 (0.35)
	GDF_{BP}	3.06 (0.12)	4.08 (0.14)	4.94 (0.14)	5.42 (0.15)	6.93 (0.20)

Note 1: The mean (standard deviation) of GDF from 500 iterations are computed. *Note* 2: GDF is computed by Monto Carlo procedure with repeated time T=200. *Note* 3: GDF_{BP} is computed by Bootstrap procedure with B=500 Bootstrap samplers.

This example shed more lights on Theorem 1. In this example, we check the validity of Theorem 1 when the model fitted a priori is partly wrong, over-fitted, under-fitted, and only include noisy predictors. Since Example 1 shows no influence of the correlation coefficients among all predictors, we only give the simulation output of ρ=0.5. Simulation results for Example 2 are summarized in Table 2. The results show that Theorem 1 is correct for all three residual distributions and all cases of models fitted a priori.

Table 2. GDF of models given a priori in Example 2, $\beta 0=(2_5 0_{15})$.

β_0	ε_i	**Model**					
		$\boldsymbol{M_1}$	$\boldsymbol{M_{5,1}}$	$\boldsymbol{M_{5,2}}$	$\boldsymbol{M_{10,1}}$	$\boldsymbol{M_{10,2}}$	$\boldsymbol{M_{20}}$
GDF	DE(0,1)	1.02 (0.42)	4.98 (0.43)	5.03 (0.58)	9.97 (0.67)	9.99 (1.08)	20.02 (0.50)
	N(0, 1)	1.02 (0.65)	4.99 (0.51)	4.96 (0.97)	10.05 (0.85)	10.00 (0.58)	19.99 (0.49)
	t(3)	1.00 (0.59)	5.00 (0.27)	5.01 (0.71)	10.00 (0.68)	10.02 (0.51)	20.02 (0.44)
GDF_{BP}	DE(0, 1)	1.39 (0.08)	4.89 (0.14)	5.55 (0.16)	9.15 (0.22)	9.66 (0.22)	16.65 (0.34)
	N(0,1)	1.29 (0.08)	5.85 (0.16)	5.50 (0.16)	10.97 (0.24)	9.64 (0.22)	18.49 (0.36)
	t(3)	1.40 (0.08)	4.28 (0.15)	5.47 (0.18)	8.32 (0.20)	8.83 (0.22)	15.60 (0.41)

Note 1: The mean (standard deviation) of GDF from 500 iterations are computed from Monto Carlo procedure. *Note* 2: The DF of M_i or $M_{i,j}$ is *i*. *Note* 3: Under-fitted model M_1={1}, correct model $M_{5,1}$={1,…,5}, wrong model $M_{5,2}$={3,…,7}, over-fitted model $M_{10,1}$={1,…,10}, noise-predictor-only model $M_{10,2}$={6,…,15}, full model M_{20}={1,…,20}.

In the following Example 3, we examine the performance of GDF as a measurement of the complexity of a modeling procedure by comparing EAICR and AICR, EGCV and GCV in model selection.

Example 3

Assume that p=10, ρ=0, 0.5 or 0.8. We study two true models: (1) $\beta 0$=010 and (2) $\beta 0$=2307.

We compare model selection performances by minimizing AICR, EAICR, GCV, and EGCV, respectively. In Table 3, we report the averaged numbers of predictors (NUM), the percentages of selecting the exact true model (CFR, correctly fitted ratio), the percentages of selecting models including the true model plus some redundant predictors (OFR, over-fitted ratio). As we expect EAICR performs much better than AICR, and EGCV works much better than GCV. In particular, the percentages of selecting the exact true model by EAICR and EGCV are much higher than one by AICR and GCV, respectively. Among all four criteria, EAICR performs best in model selection.

Table 3. Application of GDF to variable selection in Example 3.

ε_i	ρ	$\beta_0=0_{10}$			$\beta_0=2_30_7$		
		NUM	**CFR (%)**	**OFR (%)**	**NUM**	**CFR (%)**	**OFR (%)**
DE(0, 1)	0.8						
	AICR	1.51	35	65	4.01	31	69
	EAICR	0.10	92	8	3.24	89	11
	GCV	2.50	17	83	4.63	15	85
	EGCV	0.88	69	31	3.55	74	26
	0.5						
	AICR	1.43	37	63	4.04	31	69
	EAICR	0.06	95	5	3.10	91	9
	GCV	2.49	16	84	4.70	14	86
	EGCV	0.82	75	25	3.61	76	24
	0						
	AICR	1.37	24	76	4.01	37	63
	EAICR	0.10	94	6	3.09	94	6
	GCV	2.37	12	88	4.62	15	85
	EGCV	0.78	78	22	3.54	82	18
N(0, 1)	0.8						
	AICR	2.20	12	88	4.43	22	78
	EAICR	0.61	74	26	3.62	77	23
	GCV	2.86	4	96	4.92	13	87
	EGCV	1.58	53	47	4.35	59	41
	0.5						
	AICR	2.15	11	89	4.46	21	79
	EAICR	0.57	77	23	3.55	79	21
	GCV	2.83	4	96	4.97	12	88
	EGCV	1.72	58	42	4.25	62	38
	0						
	AICR	2.04	10	90	4.39	19	81
	EAICR	0.49	78	22	3.54	78	22
	GCV	2.78	2	98	4.82	11	89
	EGCV	1.48	62	38	4.21	63	37
t(3)	0.8						
	AICR	1.04	49	51	3.68	55	45
	EAICR	0.09	92	8	3.11	94	6
	GCV	2.33	17	83	4.59	22	78
	EGCV	0.86	73	27	3.66	77	23
	0.5						
	AICR	0.97	44	56	3.72	53	47
	EAICR	0.08	94	6	3.10	94	6
	GCV	2.35	13	87	4.57	22	78
	EGCV	0.78	78	22	3.57	81	19
	0						
	AICR	1.00	47	53	3.72	50	50
	EAICR	0.11	94	6	3.16	93	7

	GCV	2.25	11	89	4.61	23	77
	EGCV	0.81	80	20	3.57	83	17

Note 1: NUM is the averaged number of predictors in the selected model. *Note* 2: CFR is the percentage of selecting the exact true model M_0. *Note* 3: OFR is the percentage of over-fitted models by one additional predictor.

7. Discussion

Under the L_2 loss function and Normal assumption, Stein (1981) developed his Lemma 1 based on which SURE estimator was proposed, Ye (1998) developed his concept of the GDF, and Shen and Ye (2002) proposed their version of adaptive model selection. In this manuscript, we obtain these three parallel results under the L_1 loss function and provide two definitions of the GDF and justify their connections under the double exponential assumption. Following the steps of Ye (1998), we emphasize that we should realize the crucial difference between a model given a priori and a modeling procedure. Since a modeling procedure is the combination of both model selection and model fitting, the GDF works better than the DF while being used to adjust the bias in model selection. As a natural generalization, we also extend those results to M-estimation.

Acknowledgments

We thank the associate editor and the anonymous referee for constructive comments that substantially improved the presentation of the paper.

Appendix A.

Proof of Theorem 1

We aim to prove that (1) Nx is a finite collection of hyperplanes in Rn and (2) β^=β^(y) is a continuous function of **y** if β^ is unique.

If a LAD estimate β^ implies |Ey|=k>d, then there are k observations such that yij=x′ijβ^ for j=1,…,k. Stack them in matrix form, we have yk=Xkβ^, where yk is a k-vector and $\mathbf{X}_k$ is a $k\times d$ matrix. Let β^k be $\mathbf{X}_k^+\mathbf{y}_k$, where $\mathbf{X}_k^+$ is the unique Moore–Penrose generalized inverse of $\mathbf{X}_k$. Then $\mathbf{y}_k=\mathbf{X}_k\mathbf{X}_k^+\mathbf{y}_k$. Since rank(XkXk+)<k, **y** is in a hyperplane in Rn. There are at most ∑k=d+1n(nk) such hyperplanes. Thus (1) holds.

For any fixed $\mathbf{y}^0$, let $\mathbf{y}^m$ be a sequence converging to $\mathbf{y}^0$. We want to show that β^(ym)converges to β^(y0). Since β^(ym) is bounded we only need to show that every converging subsequence converges to β^(y0). Suppose that subsequence β^(ymk)→β˜.

$$\sum|yi0-x'i\hat\beta(y0)|=\sum|yimk-x'i\hat\beta(y0)|+\sum\{|yi0-x'i\hat\beta(y0)|-|yimk-x'i\hat\beta(y0)|\}\geq\sum|yimk-x'i\hat\beta(ymk)|+\sum\{|yi0-x'i\hat\beta(y0)|-|yimk-x'i\hat\beta(y0)|\}=\sum|yi0-x'i\hat\beta(ymk)|+Ak,$$

where $Ak=\sum\{|yi0-x'i\hat\beta(y0)|-|yimk-x'i\hat\beta(y0)|\}+\sum\{|yimk-x'i\hat\beta(ymk)|-|yi0-x'i\hat\beta(ymk)|\}$. Noting that $Ak<2\sum|yimk-yi0|\to0$, we have $\sum|yi0-x'i\hat\beta(y0)|\geq\sum|yi0-x'i\tilde\beta|$. We have β˜=β^(y0) since β^(y0) is the unique minimizer of ∑|yi0−x′iγ|. Thus (2) holds.

For any fixed $\backslash$ y0∈Rn $\backslash$ Nx, let E0 and E0c denote the elbow set based on $\mathbf{y}^0$ and its complement. From (2), we can choose small enough ε>0 such that

for \y∈Ball(y0,ε)⊂Rn\Nx we have yi−x′iβ^(y)≠0 if i∈Eyc. Therefore, E0c⊆Eyc and E0⊇Ey. From (1), |Ey|=|E0|=d, and thus Ey is locally constant with respect to **y**. This implies ∑i=1n∂μ^i/∂yi=1 if i∈Ey and ∑i=1n∂μ^i/∂yi=0 if i∈Eyc. Thus Theorem 1 is obtained. □

Proof of Theorem 2

By expansion (for simplicity, one may think about Taylor expansion in the sense of Phillips, 1991) at μi, for new data y_i^0 generated from the same mechanism generating y_i, we have

$$\sum_{i=1}^{n}|y_i^0-\hat{\mu}_i|=\sum_{i=1}^{n}|y_i^0-\mu_i|+\sum_{i=1}^{n}\operatorname{sgn}(y_i^0-\mu_i)(\mu_i-\hat{\mu}_i)+f(0)\sum_{i=1}^{n}(\mu_i-\hat{\mu}_i)^2+o_P(1),$$

$$\sum_{i=1}^{n}|y_i-\hat{\mu}_i|=\sum_{i=1}^{n}|y_i-\mu_i|+\sum_{i=1}^{n}\operatorname{sgn}(y_i-\mu_i)(\mu_i-\hat{\mu}_i)+f(0)E\sum_{i=1}^{n}(\mu_i-\hat{\mu}_i)^2+o_P(1),$$

where $f(0)$ is the density of εi at zero. Thus (9) holds because $E\{\Sigma_{i=1}^{n}|y_i^0-\mu_i|\}=E\{\Sigma_{i=1}^{n}|yi-\mu_i|\}$ and $E\{\Sigma_{i=1}^{n}\operatorname{sgn}(y_i^0-\mu_i)(\mu^i-\hat{\mu}_i)\}=0$. □

Proof of Lemma 2

$$|\tilde{y}_i-\mathbf{x}'_{\alpha i}\hat{\boldsymbol{\beta}}^{[i]}|+\sum_{j\neq i}|y_j-\mathbf{x}'_{\alpha j}\hat{\boldsymbol{\beta}}^{[i]}|=\sum_{j\neq i}|y_j-\mathbf{x}'_{\alpha j}\hat{\boldsymbol{\beta}}^{[i]}|<\sum_{j\neq i}|y_j-\mathbf{x}'_{\alpha j}\boldsymbol{\gamma}_\alpha|\le|\tilde{y}_i-\mathbf{x}'_{\alpha i}\boldsymbol{\gamma}_\alpha|+\sum_{j\neq i}|y_j-\mathbf{x}'_{\alpha j}\boldsymbol{\gamma}_\alpha|.$$

Proof of Lemma 3

If Y has generalized $t_{p,q}(0,1)$, then by integration by parts,

$$\begin{aligned}E[g'(Y)]&=g(y)c_{p,q}(1+|y|^p/q)^{-q-1/p}|_{-\infty}^{+\infty}-\int_{-\infty}^{+\infty}g(y)c_{p,q}(1+|y|^p/q)^{-q-1/p})'\,dy\\&=\int_{-\infty}^{+\infty}g(y)c_{p,q}(1+|y|^p/q)^{-q-1/p})(q+1/p)(1+|y|^p/q)(pq^{-1}|y|^{p-1})\operatorname{sgn}(y)\,dy\\&=E[g(Y)(q+1/p)(1+|y|^p/q)(pq^{-1}|y|^{p-1})\operatorname{sgn}(Y)].\end{aligned}$$

Thus we have E[g′(Y)]=E[g(Y)ψ(Y)] with ψ(u)=(q+1/p)(|u|$^{p-1}$)(1+|u|^p/q)sgn(u).

If Y follows generalized $t_{p,q}(\mu,\sigma)$, then $\tilde{Y}=(Y-\mu)/\sigma$ follows $t_{p,q}(0,1)$. Let $\tilde{g}(y)=g(\mu+\sigma y)$. Then

$E[g'(Y)]=\frac{1}{\sigma}E[\tilde{g}'(\tilde{Y})]=\frac{1}{\sigma}E[\psi(\tilde{Y})\,\tilde{g}(\tilde{Y})]=\frac{1}{\sigma}E[\psi(\frac{Y-\mu}{\sigma})g(Y)]$.

Thus Lemma 3 is shown. □

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory, pp. 267–281.

P. Craven, G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross–validation, Numerische Mathematik, 31 (1979), pp. 377-403

B. Efron. Journal of American Statistical Association, 81 (1986), pp. 461-470

B. Efron. The estimation of prediction error: covariance penalties and cross-validation, Journal of American Statistical Association, 99 (2004), pp. 619-632

B. Efron, R.J. Tibshirani. An Introduction to the Boostrap, Chapman & Hall, CRC (1993)

E.J. Hannan, B.G. Quinn. The determination of the order of autoregression, Journal of the Royal Statistical Society, Series B, 41 (1979), pp. 190-195

C.M. Hurvich, C.L. Tsai. Regression and time series model selection in small samples Biometrika, 76 (1989), pp. 297-307

Y. Li, Y. Liu, J. Zhu. Quantile regression in reproducing kernel Hilbert spaces, Journal of American Statistical Association, 102 (2007), pp. 255-268

Y. Li, J. Zhu. L_1-norm quantile regression, Journal of Computational and Graphical Statistics, 17 (2008), pp. 163-185

D. Nychka, G. Gray, P. Haaland, D. Martin, M. O'Connell. A nonparametric regression approach to syringe grading for quality improvement, Journal of American Statistical Association, 90 (1995), pp. 1171-1178

P.C.B. Phillips. A shortcut to LAD estimator asymptotics, Econometric Theory, 7 (1991), pp. 450-463

D. Pollard. Asymptotics for least absolute deviation regression estimations, Econometric Theory, 7 (1991), pp. 186-199

E. Ronchetti. Robust model selection in regression, Statistics & Probability letters, 3 (1985), pp. 21-23

G. Schwarz. Estimating the dimension of a model, The Annals of Statistics, 6 (1978), pp. 461-464

X. Shen, H. Huang, J. Ye. Adaptive model selection and assessment for exponential family distributions, Technometrics, 46 (2004), pp. 306-317

X. Shen, J. Ye. Adaptive model selection, Journal of American Statistical Association, 97 (2002), pp. 210-221

C. Stein. Estimation of the mean of a multivariate normal distribution, The Annals of Statistics, 9 (1981), pp. 1135-1151

M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society, Series B, 39 (1977), pp. 44-47

J. Ye. On measuring and correcting the effects of data mining and model selection, Journal of American Statistical Association, 93 (1998), pp. 120-131

M. Yuan. GACV for quantile smoothing splines, Computational Statistics and Data Analysis, 5 (2006), pp. 813-829